CARRA Working Paper Series

Working Paper #2014-11

# Person Matching in Historical Files using the Census Bureau's Person Validation System

Catherine G. Massey
Center for Administrative Records Research & Applications
U.S. Census Bureau
Catherine.G.Massey@Census.gov

Amy O'Hara
Center for Administrative Records Research & Applications
U.S. Census Bureau
Amy.B.Ohara@Census.gov

September 2014

Person Matching in Historical Files using the Census Bureau's Person Validation System

Catherine Massey and Amy O'Hara

September 15, 2014

Center for Administrative Records Research & Applications
U.S. Census Bureau
Catherine.G.Massey@Census.gov

Abstract

The recent release of the 1940 Census manuscripts enables the creation of longitudinal data spanning the whole of the twentieth century. Linked historical and contemporary data would allow unprecedented analyses of the causes and consequences of health, demographic, and economic change. The Census Bureau is uniquely equipped to provide high quality linkages of person records across datasets. This paper summarizes the linkage techniques employed by the Census Bureau and discusses utilization of these techniques to append protected identification keys to the 1940 Census.

Contents

## 1. Introduction

The U.S. Census Bureau is investigating the linkage of historical census data to contemporary data to facilitate research on social phenomena over the course of the 20th century. To that end, the Census Bureau is collaborating with the academic community and the Minnesota Population Center to establish the Core Longitudinal Infrastructure Population Project (CLIPP). One of the primary goals of CLIPP is to append Protected Identification Keys (PIKs) to the recently released 1940 Census manuscript using the Census Bureau's Person Identification Validation System (PVS). Once assigned PIKs, person records in the 1940 Census will be linkable to other data processed by the Census Bureau's record linkage software, such as the 2000 Census, the American Community Survey, and the Annual Social and Economic Supplement of the Current Population Survey.

CLIPP represents one of the largest, collaborative record linkage endeavors to date. The Census Bureau's Center for Administrative Records Research and Applications (CARRA) will conduct the record linkage using the Person Identification Validation System (PVS). Due to the unique challenges associated with linking historical records, the CLIPP team consists of experts in historical record linkage, as well as record linkage specialists from CARRA, who will conduct new research to improve techniques. The CLIPP project will provide an unequalled source of longitudinal data useful for health, demographic, and economic research.

This paper provides an overview of the record linkage methods employed by CARRA and a discussion of techniques to append PIKs to the 1940 Census. The next section briefly introduces the PVS, which CARRA uses to assign unique PIKs to observations in survey, federal, state, and commercial data. Section 3 discusses similarities and differences between the PVS linkage program and record linkage techniques used in the history literature. Section 4 discusses the application of an adapted version of PVS on a small sample of the 1960 Census. Sections 5 through 7 discuss strategies to assign PIKS to 1940 Census data, and Section 8 describes data quality in the PVS reference file. The final section describes future research.

## 2. PVS Overview

CARRA uses the PVS to assign unique PIKs to person records to facilitate unduplication and record linkage. The PVS uses a probabilistic matching algorithm (Felligi and Sunter, 1969) that processes data through modules, with each module blocking the data in different ways and comparing different fields to find matches. PVS modules include: Verification,[1] GeoSearch, Movers, Name, Date of Birth (DOB), and Household Composition. These modules match data to a reference file based on the Social Security Administration (SSA) Numident file (Wagner and Layne, 2014) using different combinations of personally identifiable information (PII) such as Social Security Numbers (SSNs), name, date of birth, and address. Records cascade through the modules, and only those observations that did not receive a PIK pass from one module to the next. All reference file records are available for linkage in each module.

[1] The Verification module is used when data contain reported SSN. An exact match to the Numident is performed by SSN, then the name and DOB are compared. If the name and DOB agree sufficiently, the input record is flagged as verified and the records do not cascade to any other search module. Note that censuses did not collect SSN, so none of the decennial records were processed through this module.

The PVS appends PIKs to census, survey, and administrative records data. The success of the process depends on the completeness and accuracy of the person identifiers on the input files. This paper focuses on modifications made to the production version of PVS to address the limited person identifiers in historic census data.[2]

A modified PVS uses the Name Search module to link historic census records to the Numident using full name, full date of birth (as available), and sex.[3] PVS follows the typical steps in record linkage: preprocessing the data, sorting into blocks, identifying potential matches, and resolving best matches.

First, preprocessing standardizes the blocking and matching fields between the census file and the reference file, ensuring that similar variables align.

Second, the input and reference files are sorted into blocks. Blocking creates reasonably sized search spaces to find candidate matches, which is important with large files (Michelson and Knoblock, 2006). The 2014 Numident contains nearly 500 million SSNs and forms the reference file used in the matching process. The reference file also includes all transactions associated with each SSN.[4] Consequently, the reference file is large and it is technically infeasible to compare every Numident record to every census record simultaneously.

The blocking scheme for the Name Search module uses the first letter of the first and last name fields. Using a fictitious example, Alex Aron is sorted into the A-A cut and Alex Bron is sorted into the A-B cut in both the census data and Numident data. The Name Search module compares records in the A-A census cuts to records in the A-A reference file cuts, and records in the A-B census cuts to records in the A-B reference file cuts, employing parallel processing to identify potential matches from the large reference file.

Although blocking greatly reduces the time and memory necessary to PIK data, it may introduce bias into the matched sample. For instance, similar records in different cuts may not be compared to seek a match if there is a typo in the first letter of the first or last name in the input data.[5]

Nicknames pose similar problems. The historic census data may contain a nickname where the first character is different from the first character of the formal name (or standard name). In this case, the preprocessing step of the Name Search module outputs two records for these observations, one record for the nickname and one record for the formal name. For example, if

[2] When we refer to the "production PVS," we are describing the official PVS process used by the Census Applications Branch of CARRA to append PIKs to data.

[3] Historic census data lack matching fields to apply the other PVS modules. The Verification module requires SSN, which were not collected. The Geosearch and Movers modules require name-address pairs from the same time period. There are no administrative records with address data from the early periods available. The DOB module required month-day-year data that are lacking in historic census data. The Household Composition module relies on known relationships in the census and reference file. Research is underway to incorporate parent-child links in the Name Search process for historic data.

[4] Transactions occur when corrections or name changes are made on a record of a particular SSN. The average number of claims per person is 2.1 (Harris, 2014).

[5] The formal PVS process employs several different search modules to alleviate this issue by using different blocking strategies across each module (by zipcode or date of birth, for example). Therefore, if an observation fails one module due to blocking, it may find a match in another module.

the input record has the name "Bill Smith," the formatting program will add a formal name "William" to that record. This record will then output to both the B-S cut and to the W-S cut.

Third, PVS compares census and reference file records within several passes of each block. Each pass alters the comparison of characteristics specified by the researcher, using slight variations of the match fields and permitting more fuzziness in the match. For instance, the first pass may require that first name, last name, and year of birth match exactly, but the next pass may only require year of birth to match exactly.

Within each pass of the Name Search module, potential matches are assigned a total score depending on the similarity of the characteristics of the input records and reference records. The Name Search module employs a string comparator program to measure Jaro-Winkler distances between first and last names in the input and reference files (Winkler, 1995).[6] These distances serve as a metric of how closely two names match, while allowing for some degree of misspelling.[7] For numeric variables, such as year of birth, a maximum acceptable difference between the variable value in the input and reference record is programmable. This also allows for creation of an interval, or band, around year of birth to permit inexact matches.

Potential matches are identified within each pass of the Name Search module, and only those with an overall score greater than the predetermined cutoff score are kept. Census records that do not receive a match in the first pass move to the next pass. Once the input data has been processed through all passes, potential matches are grouped into one file and sorted by person and by score.

The final step of the Name Search evaluates the potential matches. The matches with the highest scores are processed using a decision rule to determine if a PIK will be assigned. If one potential match has a higher score than all the other potential matches for a particular input record, then the PIK associated with that reference record is assigned to that input observation. If there are multiple potential matches for a particular input observation with the same high score, then no PIK is assigned.

Appending PIKs to the 1940 Census will require incorporating the techniques employed by historians into PVS due to the limited amount of PII available in historical census files. The next section discusses the techniques used by historians to link historical decennial data, as well as the differences between their techniques and PVS.

[6] The PVS string comparator was developed by William Winkler and measures the distance between two strings on a scale from 0 to 900, where a distance score of 0 is given if there is no similarity between two text strings and a score of 900 is given for an exact match. The cutoff value for the string distance is set to 750 in the Name Search module.

[7] There are clear advantages to using a string comparator instead of using phonetic codes to match names across two data sources. Using Jaro-Winkler distances allows the researcher to determine the cutoff value of an acceptable distance between two strings. The total score is calculated as the sum of the agreement and disagreement weights attributed to each matching variable (Felligi and Sunter, 1969). For the comparison of text strings, a prorated value between the chosen agreement score and chosen disagreement score is given depending on the Jaro-Winkler distance between the string in the input file and reference file.

## 3. Linking Strategies in History Literature

Historians have been linking person records across decennial censuses since the 1930s. Ferrie (1996) created one of the first national linked samples by linking the 1850 and 1960 censuses and established the standard linkage technique used in the literature. This technique relies on phonetic codes of first and last names, age, and state or country of birth to link person records.

There are several key differences between the matching process of the Name Search module and the standard techniques used by historians. First, to compare names across the input and reference file, the Name Search module employs a string comparator program to measure Jaro-Winkler distances between names in the input and reference files (Winkler, 1995). The historical linking literature, in contrast, largely favors phonetic codes to match, particularly in earlier papers. More recent papers also use string comparators. For instance some compare names using the search engine built into ancestry.com (Collins and Wanamaker, 2014), the built-in SPEDIS algorithm in SAS (Long and Ferrie, 2013), or the Freely Extensible Biomedical Record Linkage (FEBRL) software used to create the IPUMS 1850-1930 linked representative samples (Ruggles, 2006 and 2011; Goeken *et al*., 2011).

Second, the Name Search module assigns scores to potential matches depending on the similarity between the two records.[8] Using Ferrie's (1996) approach, a potential match is deemed a true match if phonetic code, birthplace, and age are the same (often with a one- to ten-year band around age). Because the Name Search module assigns a score based on the similarity between match tokens in the input and reference file, the researcher can assign different weights to different tokens and then test the sensitivity of their choices.

Ultimately, these differences result in higher match rates using the Census Bureau software than those found in the history literature.[9] This largely results from not removing common names before processing the data in PVS. Instead, after scoring potential matches and identifying the highest scored potential match for an observation, all non-unique potential matches with the same score are deleted. Although arguably less conservative, this approach yields a greater amount of matches. Another reason the Census Bureau rates are higher is the use of SSA Numident data in the reference file. The Numident contains all possible names associated with a particular SSN, which allows the Name Search module to accurately PIK individuals who change their name. In addition, because the Numident is a database of all SSNs ever issued, the Census Bureau can append PIKs to individuals who are no longer alive in later censuses.

## 4. Testing and Simulation

There has been initial testing conducted at the Census Bureau to assess PVS's adaptability to append PIKs to historical data. Specifically, a sample of 1,440 observations from the 1960 Census was transcribed from microfilm and processed by the modified Name Search module. This sample contained the limited information available on the 1960 short form: first name,

[8] For the record linkage conducted for the IPUMs 1850-1930 linked samples FEBRL linkage software was employed, which also scores potential matches. Common names were also not removed and a name commonness score was created using Jaro-Winkler distances for high frequency names (by race, birthplace, and sex) (Goeken *et al.*, 2011).

[9] Match rates in recent history papers range from 3 to 29 percent.

middle initial, last name, quarter of birth and year of birth.[10] The results from this test are promising and indicate that PVS can be successfully adapted to PIK older census data such as the 1940 Census.

*4.1 Test on 1960 Census Sample*

The 1960 test used four passes in the modified Name Search module. Table 1 reports the blocked variables by pass. All passes blocked on New York State Identification and Intelligence System (NYSIIS) phonetic codes for first and last name.[11] Pass 1 also blocked on last name, first name, year of birth, and quarter of birth. This was the most restrictive pass. Pass 2 blocked on last name and year of birth. Pass 3 blocked on first name and last name. The final pass was the least restrictive and only blocked on NYSIIS code.

Table 1: Blocking Variables in Modified Name Search Passes

| Pass | Blocking variables | Matching variables |
|---|---|---|
| 1 | • NYSIIS First Name<br>• NYSIIS Last Name<br>• First Name<br>• Last Name<br>• Year of Birth<br>• Quarter of Birth | • First Name<br>• Last Name<br>• Middle Initial<br>• Sex<br>• Quarter of Birth<br>• Year of Birth |
| 2 | • NYSIIS First Name<br>• NYSIIS Last Name<br>• Last Name<br>• Year of Birth | • First Name<br>• Last Name<br>• Middle Initial<br>• Sex<br>• Quarter of Birth<br>• Year of Birth |
| 3 | • NYSIIS First Name<br>• NYSIIS Last Name<br>• First Name<br>• Last Name | • First Name<br>• Last Name<br>• Middle Initial<br>• Sex<br>• Quarter of Birth<br>• Year of Birth |
| 4 | • NYSIIS First Name<br>• NYSIIS Last Name | • First Name<br>• Last Name<br>• Middle Initial<br>• Sex<br>• Quarter of Birth<br>• Year of Birth |

[10] For a full discussion of the 1960 test, see Massey (2014a).

[11] Blocking on both name text strings and NYSIIS codes may seem redundant. However, Mill (2012) documents instances where name strings can match even if the NYSIIS codes do not match (and vice versa). Consequently, we chose to block on NYSIIS code in all passes of this test to emulate the techniques used by economic historians. Future versions of the Modified Name Search module no longer block on NYSIIS code in each pass.

The first three passes allowed a two-year band around year of birth. This band increased to five years in the last pass. The final three passes also incorporated an array function for possible variations of first names. Arrays are particularly helpful when matching on first names because PVS allows inclusion of alternate names in the array, as well as nicknames and standard versions of names.[12] For example, if first name is "Cathy" in the input data, "Catherine" can be included in the array as a potential alternate name. The matching algorithm will then be able to match "Cathy" in the input data to "Catherine" in the reference data. For the 1940 Census data, using standardized names in the array will help prevent mismatches that may result from orthographic differences in name spellings that may arise from enumerator error.

Using the Adapted PVS on these five variables alone, 70.5 percent of observations received a PIK. When PIKed using state or country of birth as an additional match variable, the PIK rate increased from 70.5 to 75.5 percent.

For children, parents' names offer additional matching keys. In the 1960 sample, 719 of the 1,440 observations indicated they were the child of the household head. These records permitted an additional match attempt using parent's first names from the short form. Parent's first and last names are also available in the Numident for the match. The original match rate for these records was 72.2 percent. The addition of parent's names as a matching token yielded an increase in the PIK rate from 72.2 to 78.3 percent for reported children of the household head. The addition of birthplace as a matching token further increased the PIK rate to 79.0 percent.

*4.2 PIK Results from 2010 Simulation of 1960 Data*

To investigate the accuracy of the PIKs assigned to the 1960 sample, the same modified Name Search module described in Table 1 was applied to the 2010 Census. The PIKs obtained through this modified Name Search module were compared to the PIKs assigned by the Production PVS process. Out of a sample of 302,103,352 individuals processed from the 2010 Census, the modified Name Search module found a match for 192,143,629 of them, or 63.6 percent.[13] Of the observations assigned a match, the modified Name Search module assigned a PIK identical to the formal PVS PIK for 178,343,078 individuals, or 92.8 percent. This exercise shows that even if the PIK rate for the 1960 data was low, the process assigned high quality PIKs. For comparison, the Production PVS assigned PIKs to 90.3 percent of the 312.5 million 2010 Census records delivered to CARRA.[14]

[12] As part of the data preprocessing step, standardized names and nicknames are appended to records using the Name Standardizer software created by the Census Bureau's Statistical Research Division. The Name Standardizer reads in a record with the name "Matt," for example, and appends the standardized name "Matthew" to that record Therefore, researchers can include standardized names as variations for the provided name in the match. See McGaughey (1994) for more information.

[13] There were 312.5 million person records in the 2010 Census delivered to CARRA. Some of these records are missing full name or date of birth information, resulting in their omission. Only a sample of 302,103,352 records was used in this simulation. The Production PVS can process records with missing name as long as date of birth and/or address is available. If date of birth is missing, Production PVS can process records as long as name and/or address are available.

[14] The Production PVS uses full address, full name, sex, and full date of birth to conduct matches. The enhanced Production PVS reprocessed the 2010 Census and appended PIKs to 90.8 percent of person records.

Running simulations on the 2010 Census data is critical to understand the potential application of the modified Name Search used to PIK the 1960 Census data. However, the 2010 Census is not ideal to try to simulate PIK rates for the 1940 Census because it lacks birthplace. To simulate PIK rates on the 1940 data, CARRA has developed a truth deck created from the 2005 Current Population Survey respondents who supplied a SSN (Massey, 2014b). This work will assess the expected quality of the 1940 PIKs, and further investigate the characteristics of those assigned incorrect PIKs (and recalibrate accordingly).

## 5. Preparing for the 1940 Census

The Minnesota Population Center and Ancestry.com collaborated to digitize the complete, publicly-released 1940 Census. The Census Bureau is now working with the Minnesota Population Center to acquire the digitized records for linkage. In this section, we describe our proposed strategies for linking the 1940 data.

### *5.1 Linking Variables*

Name, age, and birthplace are the core matching keys available to assign PIKs to the 1940 Census. The text-string comparator built in to PVS will compare first and last names. The record linkage literature typically estimates year of birth from reported age and links records using year of birth estimates. Instead, we will use full date-of-birth information available in the Numident to calculate age on April 1, 1940 (Census Day for the 1940 Census) for each record in the reference file. Then we will assign PIKs using age observed in 1940. We observe state or country of birth in both the 1940 Census and the Numident. However, some birthplaces in the Numident necessitate aggregation due to inconsistencies in the collection of birthplaces over time. For instance, some records list “Korea” as the birthplace, while others list “South Korea” or “North Korea.” As a result, we group all three responses into one. Allowances are also necessary for former countries, such as Montenegro.

There are additional linking keys available for subsets of the population. For children, parents’ names are available from roster information in the 1940 Census and are linkable to parents’ names observed in the Numident. For the subset of the population who obtained a SSN near the time of the 1940 Census, we can also use geography information to assign PIKs. Specifically, we plan to append geography information to the reference file derived from SSN area numbers and link that information to locations observed in the 1940 Census.

### *5.2 Modules*

In addition to the Name Search module, we propose creating a birthplace module and a geography-based module. First, the data will go through the birthplace module, which will block the data by state of birth and country of birth. This module will lessen bias introduced by blocking the data by initials that may result from the Name Search module. Records that fail the birthplace module could then pass to the Name Search module. In the final module, we can block by geography information, inferred from SSN area numbers, to enhance the match.

*5.3 Representativeness*

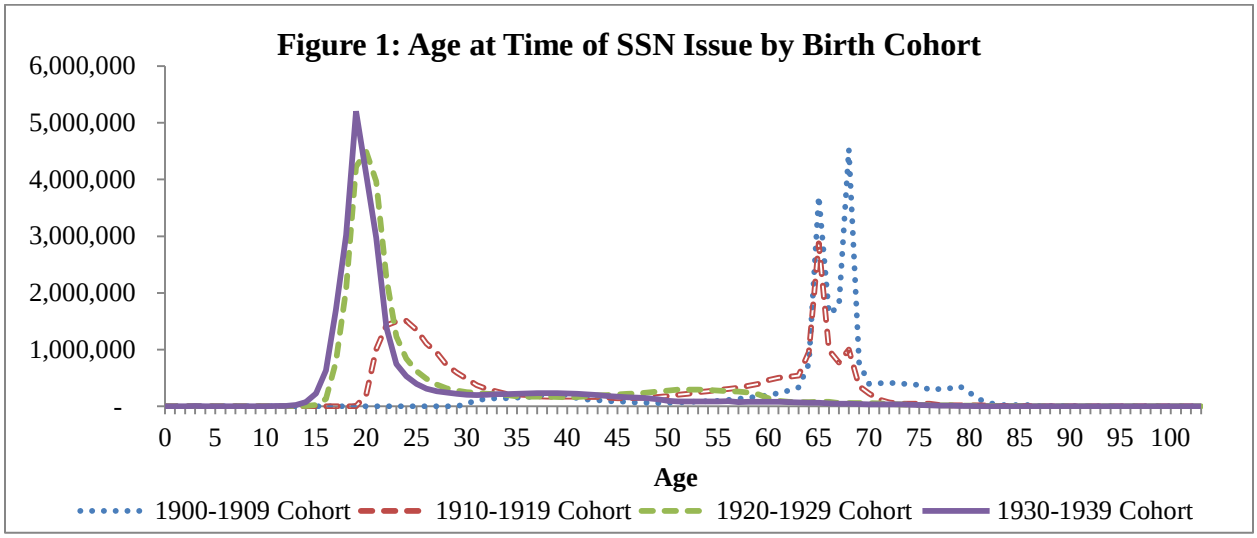
Another important consideration is the representativeness of the 1940 PIKed data. Because the reference data consists of the universe of SSNs, the most we could hope for is a PIKed sample that is representative of those alive in 1940 who also obtained a SSN. Even within the population of SSN holders, the assignment of PIKs may be nonrandom due inconsistencies in data collection methods and missing information. The method used to PIK the data may also introduce bias. For instance, the use of parent's names may result in the over representation of individuals from stable, traditional households or those with more unique names.

The Census conducts ongoing research to assess the bias in PIKed data. Bond, Brown, Luque, and O'Hara (2014), for example, found the assignment of PIKs by the formal PVS process to be nonrandom for the 2009 and 2010 American Community Surveys. They show migrants, those without health insurance, and those in poverty were less likely to be PIKed, while those in the military and the highly educated were more likely to receive a PIK. An analysis of the PIKs assigned to the 1960 data by the modified Name Search module revealed that men, wives of household heads, children, non-relatives, married, and native-born individuals were more likely to receive a PIK, where as household heads, other relatives, widows, and the foreign-born were less likely to receive a PIK (Massey, 2014a).

## 6. Incorporating SSN Area Number

Information from SSN area numbers could serve as additional information to assign PIKs. This section discusses SSN area numbers and some implementation issues to consider.

Until 1972, the first three digits of a SSN, or the area number, corresponded with the state that issued the SSN. Using this information, the CLIPP team can append geography information in the reference data and link that information to state of residence observed in the 1940 census. State of issuance is recorded in the Numident file kept by the Census. However, area-number locations will only be useful for linking those who obtained their SSN near the time we observe their location in 1940. Figure 1 charts age at SSN issuance by birth cohort.




Source: 2014 Census Numident

Figure 1 illustrates that nearly all SSN holders born before 1910 obtained their SSN when they retired (ages 59 to 65).[15] For those born between 1910 and 1920, there was a shift towards obtaining a SSN upon entrance into the work force (ages 14 to 20). The majority of those born between 1920 and 1940 obtained their SSN around age 16. Therefore, using area numbers for location on the reference side will help PIK those entering the labor force in, or around, 1940. We can also use area number information for those obtaining a SSN in 1935 and 1940 because the 1940 census asked where people lived in 1935.

## 7. Matching without Replacement

This section discusses plans to pursue matching without replacement in the 1940 PVS.

One motivation to match without replacement is to avoid assigning the same PIK to multiple observations in the 1940 Census. Currently, PVS takes all input observations in a block and then compares them to all reference observations in the same block within a module. Input records cascade through the search modules without replacement, but all reference records are available in each pass of each module.

PVS matches reference records with replacement because many input files contain duplicates. As a result, PVS was designed to enable the assignment of the same PIK to multiple input records.[16] Furthermore, the Numident may contain multiple records for each SSN. Due to the possibility of multiple records, the same SSN may be sorted into multiple blocks within a module. Once comparisons and matches are made within each of the blocks, all of the potential links are grouped together, sorted by census ID and score, and only the highest scored matches are kept.

Currently, PVS employs multiple modules, and multiple passes within each module. There are two methods we can use to incorporate matching without replacement: eliminate matched SSNs from the reference pool after each pass within a module or at the end of each module.

The first method to incorporate matching without replacement is to eliminate SSNs from the reference pool after each pass within a module. However, this approach may result in a small number of incorrect PIKs because the Numident contains multiple transactions for each SSN. Consequently, if PVS found an adequate match for a record in one pass and removed that particular record from the reference file, there is no way to know whether that reference record would have received a better match in the next pass. As a result, it may be more appropriate to eliminate matched SSNs from the reference pool at the end of a module.[17]

A second approach would eliminate matched SSNs after each module. Again, this may result in a small percentage of incorrectly assigned PIKs. If a SSN is withdrawn in the first module, there

[15] The universe of all SSN holders includes men and women. Even individuals who had never worked would obtain an SSN at age 65 if they wanted to enroll in Medicare.

[16] Often, a PIK that assigned more than once results from duplicates in the input file. For instance, the 2010 Census Match Study found 10.5 million duplicate PIKs in 2010 Census, and approximately 8.5 million of these were believed to be true duplicates resulting from households responding more than once (Rastogi and O'Hara, 2012).

[17] If we eliminate SSNs from the reference pool after each module, a PIK could still be assigned to more than one observation within a module due to the use of multiple passes.

is a chance a better match could have been found for that SSN in the next module. To avoid introducing bias from choices about the order modules, cuts, and blocking variables, we may want to address duplicates after all the data have been processed by all the modules.[18]

## 8. Variable Coverage in the Numident

The effectiveness of PVS is reliant on the quality of the input and reference data. In this section, we describe several variables in the Numident that are important for the linkage.

### *8.1 Year of Birth*

Most persons in the Numident born before 1941 have a complete date of birth. Few people have 01/01 as their birthday, which may suggest a minimal amount of heaping. Year of birth is missing for a small number of observations (2,716,107, or 1.84% of those born before 1941). Approximately 3.5 million observations are missing either month of birth or day of birth.

### *8.2 Parent's Names and Place of Birth*

Preliminary research with a small sample of the 1960 Census found that using parent's names and birthplace as additional linking variables increases the PIK rate (Massey, 2014a). Additional research is being conducted on the effect of using parent's names and birthplace on the quality of the match, as well as the extent to which they introduce bias into the matched sample. Table 2 reports the number of SSN holders born before 1941 with parent's names and birthplace recorded in the Numident.

Table 2: Sample of Individuals in Numident Born Before 1941

| | Full | % | Adults | % | Children | % |
|---|---|---|---|---|---|---|
| Father's Name | 94,723,598 | 64.30 | 47,620,118 | 48.69 | 47,103,480 | 94.16 |
| Mother's Name | 93,591,493 | 63.53 | 46,487,018 | 47.53 | 47,104,475 | 95.13 |
| N | 147,312,442 | 100.00 | 97,796,231 | 100.00 | 49,516,211 | 100.00 |

Notes: To create this table the Numident was limited to individuals born before 1941. Individuals were categorized as children if they were born in or after 1925.
Source: 2014 Census Numident

Of SSN holders born between 1925 and 1941, 94.2 percent and 95.1 percent have a father's name and mother's name recorded in the Numident, respectively. The Numident contains father's name for 64.3 percent and mother's name for 63.5 percent of SSN holders born before 1925. Those born before 1925 are less likely to reside with their parents in 1940, thus parent's names will be less helpful in the assignment of PIKs.

State of birth is reported for nearly all adults and children born before 1941. State or country of birth is missing for 7,301 observations. For 6,406 of these, city of birth is reported, which can potentially be used to find state or country of birth.

[18] Future research is needed to determine if eliminating SSNs after each pass, or after each module, results in lower error rates.

*8.3 SSN Issue Dates*

Accurate information on SSN issue dates is necessary to use SSN area numbers in matching. Date of issuance is missing or incomplete for 14,540,298 SSN holders born before 1941. However, we may be able to use group number information (the fourth and fifth digits of an SSN) to determine date of issuance for these cases.

*8.4 Historical Records in Numident*

SSNs were first issued in 1936. Initial confusion surrounding their purpose may create additional problems for appending PIKs to older data. According to the SSA website, possibly three to four percent of SSNs were issued to individuals who had already obtained a SSN. These duplicates resulted from individuals applying for new SSNs if they misplaced their card or obtained a new job (Corson, 1938). Because each SSN is associated with a different PIK, individuals with multiple PIKs in the reference file will be difficult to accurately PIK in the 1940 Census. Research by National Opinion Research Center (NORC) estimates a lower bound of 0.3 percent for duplicates in the entire reference file (NORC, 2013); however, the percentage may be higher for those alive in 1940.

## 9. Additional Research Needed

Technical documentation is necessary to describe the preprocessing steps for the 1940 Census and the Numident. In particular, we need to document how place of birth is coded and the use of alternate names. The linked 1940 data will also need to be compatible with contemporary census data (e.g., birthplace codes) for analyses linking the 1940 Census to contemporary data.

Research is important to understand the uniqueness of name/date-of-birth/place-of-birth combinations in the Numident, run sensitivity analyses, and conduct simulations to understand potential error rates. In addition, research is necessary to understand SSN assignment and area numbers. The quality of historical SSA data is critical to the success of appending PIKs to older data, and we need a full understanding of area numbers to integrate location information into PVS. Such research will help us understand the quality of the match.

Once the data has PIKs, the team will create documentation for the data, and produce reports that assess the representativeness of the PIKed 1940 data, as well as the representativeness of the 1940-2000 linked data.

10. References


Bond, B., Brown, J., Luque, A. & O'Hara, A., 2014. The Nature of Bias When Studying Only Linked Person Records: Evidence from the American Community Survey. *CARRA Working Paper #2013-08.*

Collins, W. J. & Wanamaker, M. H., 2014. Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics,* 6(1), pp. 220-52.

Corson, J. J., 1938. Administering Old-Age Insurance. *Social Security Bulletin,* 1(5), pp. 3-6.

Fellegi, I. P. & Sunter, A. B., 1969. A Theory for Record Linkage. *Journal of the American Statistical Association,* Volume 64, pp. 1183-1210.

Ferrie, J., 1996. A New Sample of Americans Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules. *Historical Methods,* Volume 34, pp. 141-56.

Goeken, R., Huynh, L., Lynch, T. A. & Vick, R., 2011. New Methods of Census Record Linking. *Historical Methods,* 44(1), pp. 7-14.

Harris, B., 2014. Transgender Labor Supply, Employment, and Earnings Gaps: Evidence from the Federal Administrative Records and the American Community Survey. *CARRA Working Paper.*

Long, J. & Ferrie, J., 2013. Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Review,* 103(4), pp. 1109-37.

Massey, C. G., 2014a. Creating Linked Historical Data: An Assessment of the Census Bureau's Ability to Assign Protected Identification Keys to the 1960 Census. *CARRA Working Paper 2014-12.*

Massey, C. G., 2014b. Playing with Matches: An Assessment of Match Accuracy in Linked Historical Data. *CARRA Working Paper 2014-XX.*

McGaughey, A., 1994. The 1995 Bureau of the Census Computer Name Standardizer Documentation. *Statistical Research Division Research Paper.*

Michelson, M. & Knoblock, C. A., 2006. Learning Blocking Schemes for Record Linkage. *Proceedings of the 21st National Conference on Artificial Intelligence,* Volume AAAI-06.

Mill, R., 2012. Assessing Individual-Level Record Linkage between Historical Datasets. *Preliminary Working Paper.*

NORC, 2013. PVS Research: Task 4, Further PVS Research Final Research Report.

Rastogi, S. & O'Hara, A., 2012. *2010 Census Match Study,* Washington, DC: United States Department of Commerce.

Ruggles, S., 2006. . Linking historical censuses: A new approach. *History and Computing,* Volume 14, pp. 213-24.

Ruggles, S., 2011. . Intergenerational coresidence and family transitions in the united states,. *Journal of Marriage and Family,* 73(1), p. :136–148.

Wagner, D. & Layne, M., 2014. The Person Identification Validation System: Applying the Center for Administrative Records and Research and Applications' Record Linkage Software. *Center for Administrative Records Research and Applications Report Series (#2014-01).*

Winkler, W. E., 1995. Matching and Record Linkage. In: *Business Survey Methods.* New York: J. Wiley, pp. 355-384.